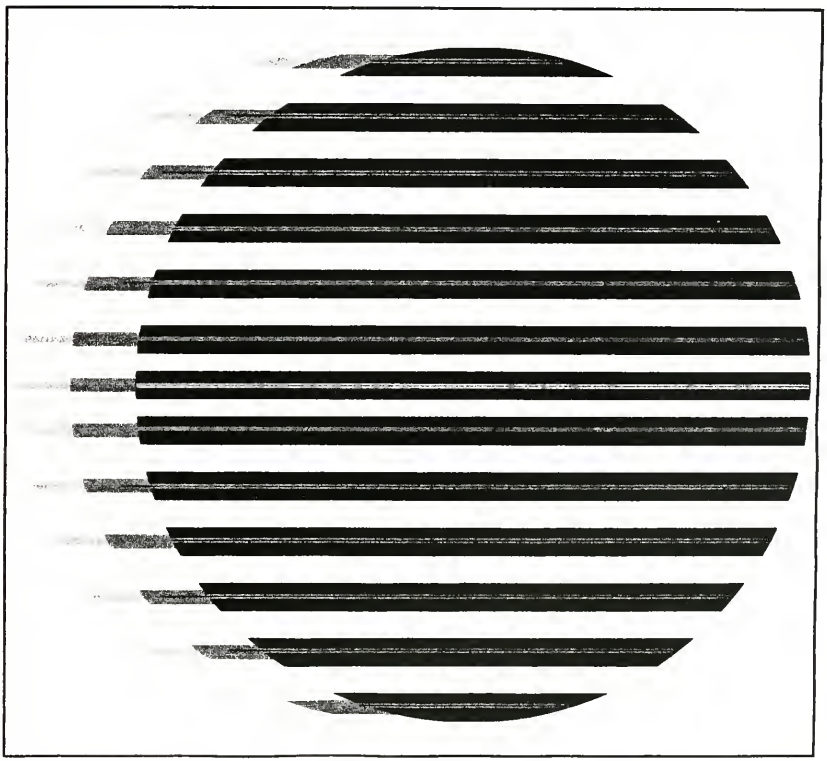
IASSIST

QUARTERLY

VOLUME 20 Winter 1996 NUMBER 4



Printed in the U.S.A.

IASSIST QUARTERLY

The **IASSIST QUARTERLY** represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The **QUARTERLY** reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of **IASSIST**.

Information for Authors

The **QUARTERLY** is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press *Manual of Style* or Kate L. Turabian's *Manual for Writers*. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". *Journal of the American Society for Information Science 30(2):77-82, March 1979*. If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor: Laura Bartolo, Libraries & Media Services, Kent State University, Kent, Ohio 44242 (216) 672-3024 Email: LBARTOLO@KENTVM.KENT.EDU Book reviews should be submitted in duplicate to the Book Review Editor: Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California 92713 USA. (714) 856-4978 E-Mail: DTSANG@ORION.CF.UCI.ED

Title: **Newsletter - International Association for Social Science Information Service and Technology**
ISSN - United States: 0739-1137

CONTENTS

Volume 20 **Number 4** **Winter 1996**

FEATURES

Facilitating Access to Comparative Data

by Ekkehard Mochmann & Lorenz Gräf,
Central Archive for Empirical Social Research (ZA);
at the University of Cologne

Mandate of the ZA as part of a Social Science Infrastructure
The Central Archive for Empirical Social Research (ZA) at the University of Cologne serves as a research, training and resource center for social research. Founded in 1960 by the Faculty of Economics and Social Sciences of the University of Cologne, it soon developed into a data service with a supraregional and international clientele. As a central node in the international data service network it became the starting point for a more comprehensive social science infrastructure, the German Social Science Infrastructure Services (GESIS e. V.). This association was created as a response to needs formulated by the social science profession in 1986 to provide infrastructural services in all fields of social research with particular emphasis on:

- collecting data and making it available for further research.
- informing about social science literature and research projects.
- development of research methods, teaching instruments and methods consulting for research projects.

The core of the ZA mandate is to facilitate access to already existing data, especially survey data, which can be used for secondary analysis. The holdings cover all fields of empirical social research. Beyond survey data there are collections of statistical data, regional data, various types of quantitative historical data and machine readable texts for computer assisted content analysis, as well as party manifestos and other text collections.

ZA provides services in the area of acquisition, processing, documenting and making available data for social research, especially survey data. ZA offers consulting services for secondary analysis. Training in complex analysis methods takes place twice a year in the ZA spring seminar for empirical social research and the autumn seminar for quantitative historical research. Beyond this, ZA creates ex post statistical time series and supports comparative international studies for the analysis of long term social developments .

ZA holdings of empirical social research data include European time series and comparative studies. The ZA department ZHSF (Center for Historical Social Research) develops data bases, in some cases going back to earlier centuries. The ZA holds nearly 4000 data sets and data collections. Even though there is no particular topical restriction, emphasis is on topics such as political attitudes, election studies, education, unemployment, leisure and occupation, media and the environment.

Among the data sets intensively used are the EUROBAROMETERs (a data pool of comparative surveys from European countries taken for more than 15 years), the German General Social Survey ALLBUS, which is conducted every two years, the International Social Survey Program (ISSP) for 25 countries from Australia, America, Europe to Japan. Similar attention is paid to the monthly POLITBAROMETER series provided by the Research Group Elections (FGW: Forschungsgruppe Wahlen) which is also presented on the second public TV station (ZDF) every month and the collection of surveys to the national parliament (Bundestag) since 1949.

A GESIS branch in Berlin is now focusing on data and information transfer from and to Eastern Europe. Recently more than 400 data sets from surveys conducted in the former German Democratic Republic (GDR) since 1975, were included in the ZA holdings and were processed for secondary analyses. Currently emphasis is on supporting initiatives to create infrastructure institutes in Eastern Europe and to develop a service network for European wide data transfer.

The ZA has access to data held in the social science data archives world wide. International data transfer is coordinated with the Council of European Social Science Data Archives (CESSDA) and the International Federation of Data Organizations for the Social Sciences (IFDO). Access to internationally distributed data bases is supported by making use of modern telematic services like WAIS, WWW, FTP on the INTERNET and other computer networks.

Selecting relevant data and solving methodological problems relating to secondary analysis is an essential part of individual consulting. The newsletter ZA INFORMATION and the journal Historical Social Research (HSR) inform about new data

sets, methodological developments, research findings and conferences. A documentation of more than 1000 empirical research projects conducted in Germany, Austria and Switzerland is published annually.

Organizational priorities in collecting and distributing data

From the very beginning the ZA philosophy was to develop services in close interaction with the scientific community. As a consequence of this philosophy ZA also supports a small research and training department which focuses on new methodological developments in data collection and analysis. Under a guest professor scheme scholars from abroad are supported in their research from planning new surveys to secondary analysis of available data. Experts in data management and analysis offer advice from the selection of appropriate data to advanced statistical analysis.

Already in the 60s Erwin K. Scheuch, one of the ZA founding fathers, created a climate for comparative research which was inspired by the Standing Committee for Comparative Research of the International Social Science Council, in which he cooperated with Stein Rokkan and Warren Miller. This orientation was enforced by the emerging European Unification and the globalization of social research.

Over the years several international research projects have chosen the ZA as their resource center for creating an integrated data bases. Integrating national data sets into internationally comparative data sets includes comprehensive documentation of methodological, technical and historical background of a study and additional interpretation knowledge to facilitate further comparative analysis. Currently ZA serves in this function for the EUROBAROMETERS (jointly with ICPSR and Swedish Data Services (SSD), the International Social Survey Program (ISSP), and the major election studies to national parliaments in Europe (ICORE)[2]. Bringing together researchers working on the data and the data management experience of ZA provides a unique working environment for creating an integrated fund of knowledge on core topics of European social development. In cooperation with the principal investigators and other European data services ZA coordinates and creates European data bases, which could otherwise not be made available to the scientific community, relying just on national resources.

ZA strongly supports a policy of labor division between European archives according to topically focused European data collections. Under tightening resources this is a must for integrating the European data bases. Inspite of intellectual and political efforts there is an ongoing demand for additional European resources to achieve what cannot be covered by the subsidiarity principle: the data service capacities are by and large absorbed by the national demands and there is little leverage to cope with additional international workloads.

Direct access to the expertise and information banks on social science literature and research projects, as well as to the methodological expertise of its GESIS partner institutes complement this infrastructural support for the production and analysis of comparative data bases on Europe.

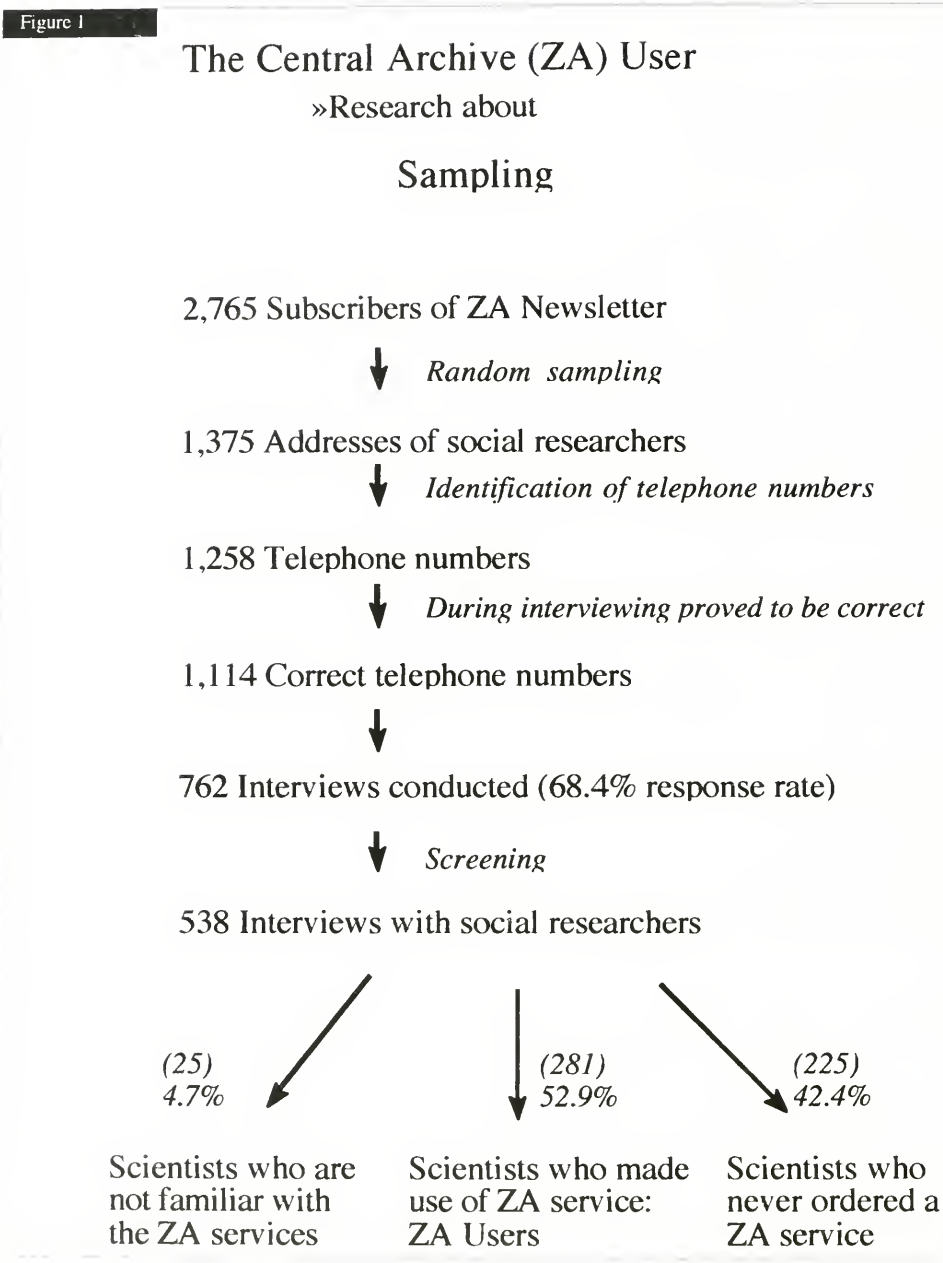
The ZA User Survey

Although the central archive has always made efforts to communicate with its clientele we have found it necessary to get more information about our clientele to face the rapid technological changes which are taking place. In the past, information concerning needs and demands of the clientele were mostly gathered by mail surveys. This procedure involves three major drawbacks. First, only the users of the institution are surveyed. So we would miss the comments of those researchers who did not make use of the data services. Second, the findings are often biased because only the most motivated people contribute in these surveys. Third, the response rates in mail surveys are low. The installation of a laboratory for telephone surveys at the University of Cologne last autumn gave us the opportunity to avoid these drawbacks. In a pilot study to test this telephone facility we could interview social researchers about their research environment and about their impression of the central archive.

Description of the sampling procedure

The target population of this study were all social scientists engaged in empirical research. For this purpose we defined empirical social research as a quantitative approach which is done with the methods of empirical social research, mainly interviewing, observation and content or document analysis (cf. Obershall 1972) Since there is no list of scientists using this very approach in their research, we could have started our project with a list of institutions known to us as informants for our documentation. But this procedure would have led to some sort of snowball sample resulting in an unpredictable sample structure. Furthermore, we wanted to interview even those people who do social research but do not want to appear in our documentation so they do not inform us about their work. Eventually, we came to the conclusion that a sample drawn out of the subscribers of the ZA Newsletter would suit our needs best, for they may be assumed to be highly interested in the application of the methods of social research. It was equally important for our purpose that fifty percent of the ZA Users were also subscribers to the Newsletter. Sampling under the subscribers of the ZA Newsletter gave us the opportunity to get the

Figure 1


The Central Archive (ZA) User
»Research about
Sampling
2,765 Subscribers of ZA Newsletter
Random sampling
1,375 Addresses of social researchers
Identification of telephone numbers
1,258 Telephone numbers
During interviewing proved to be correct
1,114 Correct telephone numbers
762 Interviews conducted (68.4% response rate)
Screening
538 Interviews with social researchers
(25)
4.7%
(281)
52.9%
(225)
42.4%
Scientists who are not familiar with the ZA services
Scientists who made use of ZA service: ZA Users
Scientists who never ordered a ZA service

feedback of those who had already made use of the services of the Central Archive, and of potential users as well.

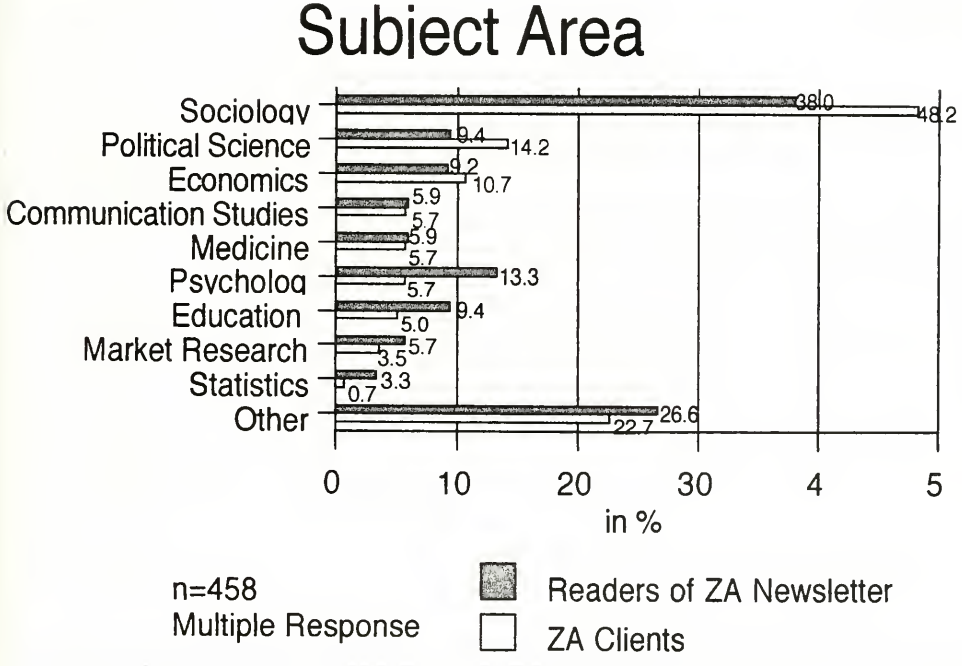
We planned to collect about 500 interviews. In advance, we estimated that about 25% of the subscribers were not directly involved in empirical research, e.g. librarians or local staff of university computer centers. From over 2,500 subscribers of the ZA Newsletter we drew a sample of 1,375 people. Since we only had their addresses we had to find out their phone numbers. Using a telephone directory CD-ROM and directory inquiries we managed to locate more than 1,100 potential respondents. The survey took place between Nov. 28 and Dec. 5, 1995. We conducted over 700 interviews. More than 200 respondents were not engaged in empirical social research so we finished with a sample of 538 social researchers. Figure 1 shows the details of the sampling procedure.

The interviews consisted of three parts. The first part dealt with the institutional affiliation of the researcher. The description of the actual empirical work formed the content of the second part and finally the respondents were asked questions concerning the performance of the Central Archive. In this paper we will focus on the description of the research community and the ZA clientele.

Characteristics of the ZA Clientele

Empirical social research is done in a variety of disciplines. The readers of the ZA Newsletter are heavily inclined to sociology as shown in figure 2. Two in five researchers (38.0%) belong to an institute which is situated in the field of sociology. The relevance of sociology is outstanding. It is mentioned nearly three times as often than is psychology (13.3%), which ranges second. Next follows a group of three fields with a proportion of ten percent each: Economics, Political Science and Education. These five subjects together form the core of the Social Sciences. Medicine, Communication Studies and

Figure 2

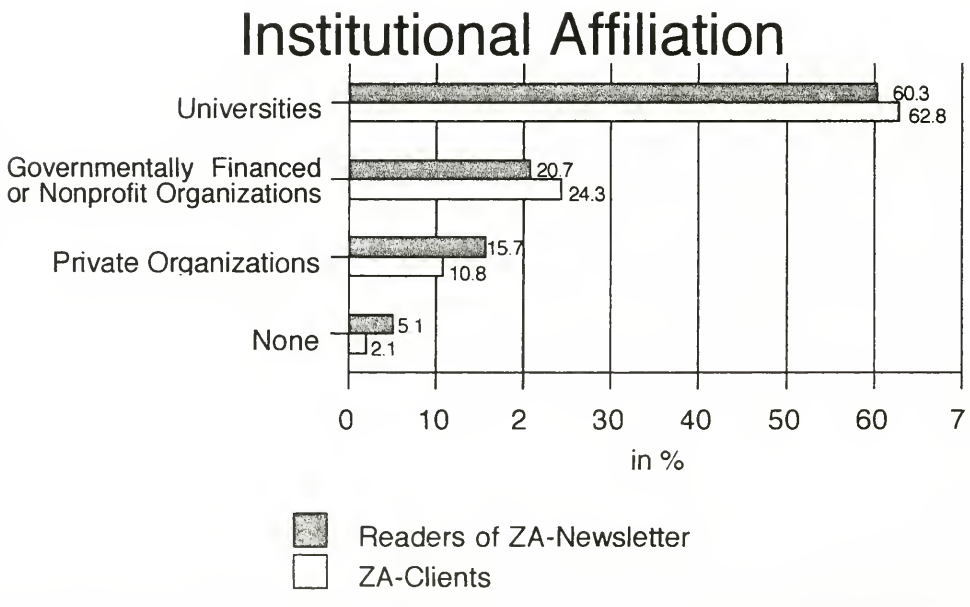


Market Research gain nearly 5% each. Statistics ranges last in this list of disciplines. There were 27 other subject fields named by the respondents, like Geography, History, Criminology, Social Psychology etc. But none gained more than two percent.

Let us now focus on those interviewees who have formerly received data sets from the Central Archive. Among those people nearly 50% belong to an institute of the research field of Sociology. The second most important discipline is Political Science. Psychology which was second among the readership of the ZA Newsletter now follows in the fourth position. This is due to the fact that psychologists do not deal that much with survey data. They prefer experimental data mostly collected from college graduates. They adopt the methods of empirical research but they normally do not need nationwide survey data. Political scientists on the other hand very often look for election data or data concerning the nationwide electorate. So it is not surprising that they range second as users of the ZA data service. Ranging third among the users of the ZA Data service are researchers belonging to institutions in the field of Economics. They gain a proportion of nearly ten percent. Psychology, Education, Communication Science and Medicine gain 5% each. Obtaining data from the Central Archive is of less importance for people belonging to Market Research Institutes and to the Statistics Branch. The former do not care much about surveys carried out by other scientists and the statisticians do not seem to be in particular demand of survey data.

As shown in figure 3 two thirds of the users of the Central Archive work in an academic institutional background. 20.7% of the respondents are employed in publicly financed research institutes. They consist mainly of federal research agencies, like the Bundesinstitut für Bildungsforschung, or governmentally financed large scale institutes, like the Max-Planck-Institute or the Wissenschaftszentrum Berlin für Sozialforschung. Only a small number of them are private nonprofit organizations like the Konrad-Adenauer-Stiftung. 13.9% of the readers of the ZA Newsletter work in private organisations in the commercial sector, mainly within the field of market research. There are only slight differences in the percentage between the researchers who have made use of the ZA data service and those who have not. While emphasis is on providing services for the academic community, the clientele also includes researchers from public administration and the media. In the ZA User Survey people

Figure 3



who work in the media are underrepresented to some extent. They normally do not subscribe to the ZA Newsletter because they are less interested in methological issues. The survey focused on people who are actually doing social research. But a remarkable part of the ZA clientele is mainly interested in getting information about the distribution of attitudes in the population.

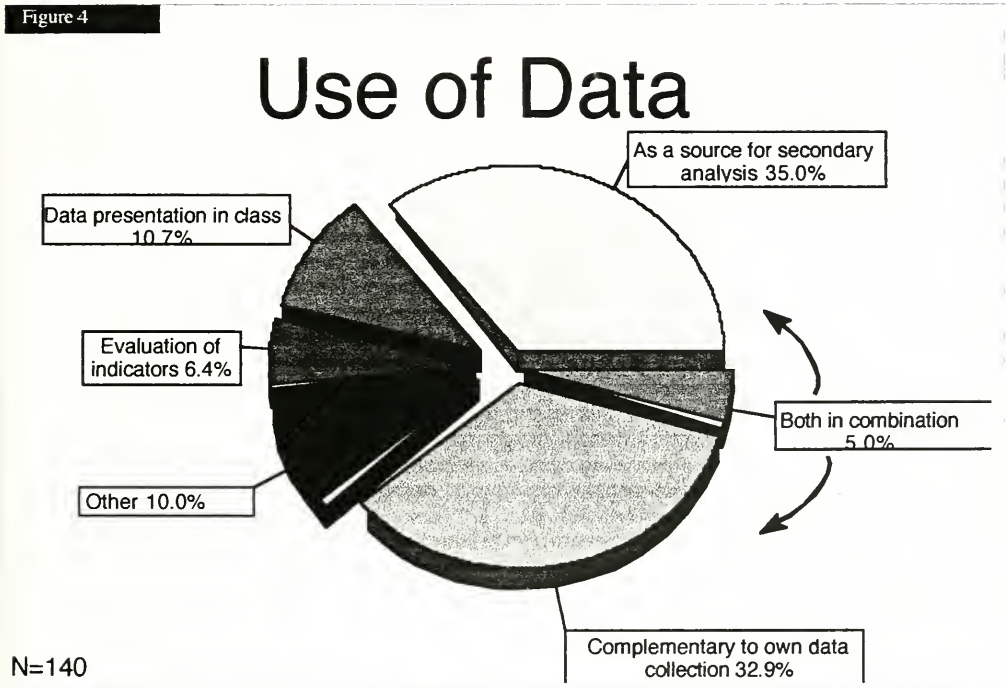
For what purpose do the clients use the data?

We asked those persons who had at least once received data from the Cologne Archive what purpose they followed in examining the data. The question was posed as open ended question and respondents could give multiple responses. Nevertheless we got a clear-cut picture. There are two main intentions behind the ordering of data. One third of the population uses the data as a source for secondary analysis under a new research question. Another third employs the data as a supplement to own data sets. This completion was mostly sought in time dimension. In most cases this means that researchers who have already got data at the present stage want to make comparisons concerning the same population at some former point in time. The intention to conduct an international or intercultural comparison as a supplement to their own data was mentioned by a smaller fraction. 10.7% of the researchers used the data for teaching purposes in class and 6.4% used the data in order to evaluate indicators used by other researchers. Another 10% named other intentions, like information about the distribution of certain opinions in society, compilation of dissertation thesis etc.

Internet as a Research Tool

In the near future the already heavy use of the internet by the research community will drastically change and hopefully improve the conditions for doing scientific research. Computer mediated communication via email will expand the possibility to collaborate and interact with distant colleagues. The flow of information will accelerate and increase when scientists begin to use virtual arenas (multi user dungeons, news groups, mailing lists, virtual conferences, electronic journals etc.) to discuss and distribute new ideas. With more scientists using the internet the demand for quick and easy access to socio-economic data will increase. Therefore it is vital for data archives to know how many of their clients have access to the internet and for

Figure 4



how many of them it has already become an ordinary research tool.

In November 1995 when we asked German social scientists 56.7% of them had direct access to the internet from their working place and 35.2% made frequent use of it. This low percentage indicates that the internet-revolution has still to gain ground in Germany. Only one third of the subscribers of the ZA Newsletter have adopted the internet as a research tool. Broken down into institutional affiliation we find a big difference between researchers working in the academic context and those who work outside university. Already 68.4% of the respondents belonging to university institutes have access to the internet. This figure is nearly twice as high as in non-academic institutes. In private organisations only 32.2% of the employees dispose of a direct connection to the internet. In governmentally financed and nonprofit institutes the adoption rate is higher and amounts to 42.7%.

	Access	Frequent use	Ratio (use/access)	N
University	68.4	44.1	64.5	(320)
Government / Nonprofit	42.7	26.1	61.1	(119)
Industry	32.2	15.1	46.9	(73)
No institutional affiliation	33.3	26.7	80.2	(15)
	56.7	35.2	62.1	(527)

Table 1: Access and use of the internet by institutional affiliation

Access to the internet does not imply that scientists make use of the internet in their daily work. Only two thirds of the researchers with access to the internet adopt an internet based service as an ordinary research tool. There is still a lot of hesitation in exploring the usefulness of the internet. The ratio of use to access of the internet is nearly the same in university institutes and in governmentally financed institutes. But in the industrial context only one half of the people with direct access to the internet make frequent use of email, WWW, FTP or some other internet service. The situation seems even worse if we look at the percentage of scientists in the industrial context using the internet. Only 15.1% of them mention the internet as a useful research tool. In Germany, at the time of our survey, the internet was still an academic challenge. But we suppose that the low adoption-rates in the industry sector are only due to the fact that the internet is basically an academic invention. With a time lag of a few months we expect that researchers in non-academic institutes will use the internet with the same frequency as their colleagues in university institutes.

It is often assumed that one of the major impacts of the use of internet-services will be a gap between young and skilled persons who adopt the new technologies quickly and older people who will be excluded from the new information technologies (cf. Negroponte 1995). In our study we do not find support for the thesis of a widening gap between the generations. We do find differences between young and older scientists in the access-rates but there are no differences in adoption-rates. Since the internet has not yet arrived in non-academic institutes we confine the analysis of this thesis to institutes in universities. As shown in table 2 nearly three quarters (73.8%) of the young scientists (age < 40) have access to the internet. Among the older scientists (age ≥ 40) 65.6% dispose of a direct access. If we focus on those people in universities who dispose of a direct access to the internet the percentage of frequent users among young scientists is nearly the same as among older scientists. 65.9% of the young scientists make frequent use of internet-services compared to 64.0% of older scientists. Thus the adoption-rates in the two age-groups are almost identical. If there was an effect of age on adoption we would expect a much higher adoption-rate among the younger scientists. We can conclude from these findings that differences in the percentage of internet-users between age-groups are only the result of differences in access-rates. Presumably older people get access to the internet at a later stage of the innovation-process than younger people. But if there is a direct access to the internet the same fraction of researchers will use the internet in the older and in the younger generation. This indicates that the use of internet-services is already a valuable research tool and that it depends mostly on the institutional context in which the scientist works whether he adopts internet-services or does not. But we can expect that in the near future the use of internet-services will be as natural as that of personal computers is now. Therefore archives have started to prepare themselves for the coming internet age.

	Access	Frequent use	Ratio (use/access)	N
under 40 years	73.8	48.6	65.9	(107)
40 years and older	65.6	42.0	64.0	(212)
only university institues				

Table 2: Access and use of the internet by age

Desiderata and Recommendations of the ZA Clientele

At the end of the interview we asked the respondents if there was anything that the Central Archive should improve or which services should be introduced. A large fraction of the respondents commented on the information policy. They wanted information to come more frequently and more directly to their working place. Another group recommended to give more detailed information. Some researchers gave the advice to foster the effort of addressing people outside the core-disciplines of Social Sciences. The second main topic was the dissemination of data. Many of the users wanted to have quick and easy access to the data via FTP and to have more data sets made available on CD-ROM. Some mentioned the present pricing policy and expressed their wish for reduced charges for data access. As the third main topic some users pointed to topically focused data collections and to a better and easier access to international data. Some researchers would be glad if we could offer more surveys from the field of commercial market research and if we could offer more recent data.

Facilitating Access to Comparative Data

Using the Internet and Publishing on CD-ROM

The central archive has always made the effort to expand its services and to use new technology to disseminate data and to communicate with its clientele as shown in the first chapter of this paper. In response to the answers the researchers gave in the user survey the central archive will strengthen these efforts. We will spread information about new data sets and other relevant news through a mailing list. More detailed and always up-to-date information can be found on our web pages (http://www.za.uni-koeln.de/) just now and will be developed further in the near future.The question text, codebook information and marginal distributions of the International Social Survey Program (ISSP) e.g. are searchable in the Internet under WAIS. Soon data will be accessible by FTP-Transfer. Furthermore we will enlarge our collections of data sets available on CD-ROM. Third we are engaged in a multinational project, named ILSES which aims at the development of an integrated library- and dataservice. Finally, we have installed a scientific laboratory equipped with all the infrastructure needed for comparative research. Also, the European Data Archives are creating a virtually integrated catalogue of their holdings, accessible via Internet.

Social Research Labs / Large Scale Facilities

As we start aging in the virtual scientific community we learn that the dream of information and data traveling to any place in the world is becoming true, yet it does not provide the ideal research environment for comparative research. Researchers may be well informed about major events in their societies that might have had an impact on attitudes and behavior of respondents. The further we progress in time, the more interpretation knowledge must be transferred to the collective memory of researchers in order to provide the context that was decisive in the phase of data collection. This is particularly relevant for information about other societies which are not part of the daily information routine of the researcher.

Contextual information, cultural background and historic knowledge which may be necessary for sound interpretation of empirical evidence do not automatically travel with the collection of data sets from different societies . Bringing together relevant data is still an exercise in systematic selection of comparable variables, data recoding and overcoming transboarder data flow hurdles emanating from data protection and data access regulations.

A response to the needs of comparative research may be social science data labs, in which all relevant data and information for a particular research field is at the fingertips.Over the past two years ZA has created a EUROLAB which provides access to major comparative studies and related background material (e.g. party manifestos, media-reports, event data bases, fact books etc.). The study collections include among others the International Social Survey Program, the Eurobarometers and major election studies on national parliaments in Europe.

The Standing Committee for the Social Sciences of the European Science Foundation had pointed to the need for better integration of the European data base and brought to the attention of the European Union that social science data bases are the equivalent to large scale research instruments of the natural and technical sciences. A study panel proposed to acknowledge

the need of social research for Large Scale facilities where researchers not normally having access, could come to profit from available resources. The Institute for Social Sciences in Essex and the Zentralarchiv in Cologne received recognition as first Large Scale Facilities in Europe under the Training and Mobility Program of the EU. This will allow to cover travel and subsistence costs for scholars from EU member and associated states who want to make use of these resources subject to approval of their applications.

Over the next three years this will allow the ZA to have scholars and research teams not only making use of the resources, but at the same time enjoying truely comparative research by bringing together their specific knowledge about different countries. Thus they can help to validate data and background material. Ultimately this will improve the research resources for the scientific community at large, since validated data and background information may be compiled in knowledge basis for general distribution.

References:

Lazarsfeld, Paul F., 1962: The Sociology of Empirical Social Research; in: American Sociological Review, Volume 27, No. 6

Lazarsfeld, Paul F., 1972: Foreword; in: Anthony Oberschall (Ed.): The Establishment of Empirical Sociology: Studies in Continuity, Discontinuity, and Institutionalization; New York

Marcson, Simon, 1972: Research Settings; in: Saad Z. Nagi and Ronald G. Corwin (Ed.): The Social Contexts of Research; London

Negroponte, Nicholas, 1995: *Being digital*, New York

Oberschall Anthony, 1972: Introduction: The Sociological Study of the History of Social Research; in: Anthony Oberschall (Ed.): The Establishment of Empirical Sociology: Studies in Continuity, Discontinuity, and Institutionalization; New York

Zuckerman, Harriet, 1988: The Sociology of Science; in: Neil J. Smelser (Ed.): Handbook of Sociology; Newbury Park, London, New Delhi

1.Paper presented at the annual meetings of IASSIST, May 15, 1996, Minneapolis, Minnesota.

2. International Community for Research into Elections and Representative Democracy

Comments on the Data Access and Dissemination System

by Lisa J. Neidert[1],
Data Archive, Population Studies Center
University of Michigan

The Data Access and Dissemination System (DADS) will be the vehicle for dissemination of census data in the year 2000. Information distributed in published census volumes in 1990 will be accessed from the internet for future censuses. Complete data files will no longer be written to media for redistribution to users. Instead, users will access DADS and pull off the tables they need. The advantages to the U.S. Census Bureau and its customers are quicker turnaround for release of files, cost-effectiveness, and increased access.

Several factors have probably motivated the Census Bureau to make this move. First, all federal agencies are responding to Al Gore's call for internet access by January 1, 1996. Second, changes in technology, such as the development of the internet, high speed computers, and low-cost storage have made this method of distribution feasible. In the past year, we have witnessed an explosion in the number of websites that distribute data (e.g. PSID, HRS/AHEAD, National Longitudinal Surveys, IPUMS, Milwaukee Parental Choice Program, Wisconsin Longitudinal Study, Russian Longitudinal Monitoring Survey, World Fertility Surveys, Malaysia Family Life Survey, Survey of Families and Households.) Finally, distributing information via DADS is a cheaper alternative for the Census Bureau, particularly when compared with the cost of printing.

As with any change, however, there are probably people who were better served by the old dissemination methods than they will be by DADS. It is clear that the Census Bureau wants this system to serve all users. However, there are some shortcomings that should be solved in upcoming renditions of DADS if the Census Bureau is to reach that goal.

The Data Access and Dissemination System doesn't exist yet. It is still a concept. However, I will use the "Data Access" page on the Census Bureau's Web site provides a good working model of DADS; and much of it is likely to be incorporated into the future operational DADS. It is also likely that many features of this current system will be remodeled, so some of my comments may be "old news" to the inner circle of DADS developers.

The current configuration of DADS needs three important improvements. First, DADS should provide the same information that one could get using the old dissemination methods. The data may be in a different form than they were in the past, but the content of the data must not be compromised. When the PSID changed from a family/individual file with a reocrd length approaching 32,767 to its new form of family records and individual records, the same information was still available. It takes new knowledge to work with the data, but users can still create the same sorts of tables they could create in the past. In contrast, the current configuration of DADS does not allow users to create all the tables they could in the past. Second, DADS should accommodate the users who have access to high-speed computers and large amounts of disk space. Some users would like to have the data on their own systems, rather than requesting tables and extracts from DADS. The FTP access of raw files is weak in the Data Access site. If FTP access to raw files proves to be impossible because of the need to protect respondent confidentiality, then the extraction system should be improved. Finally, and perhaps foremost in the minds of data librarians and archivists, there is the question of whether DADS will meet the archival needs of future users of census data. Expertise on and access to state and federal records tends to be fairly short-lived. Thus, it is essential that the archival needs of future researcheers be considered in the development of DADS. What sorts of records will be turned over to the National Archives and in what form?

Loss of Information with DADS

Most users of summary tape files (STF) find the summary-level sequence charts confusing. However,the current configuration of the Data Access system does not make it clear that all the choices available in a typical summary tape file are available in the new system. Users can get tabulations for states, counties, metropolitan statistical areas (MSAs), tracts, and blocks—the most typical choices. But can they get tabulations for central cities of MSAs (summary level 340) or for any of the American Indian Reservation categorizations (summary levels 210-221)? What about county-specific zip code statistics (summary level 820 versus summary level 810)?

The way the Data Access system is currently configured some items in the geographic identification section are not accesible. Occasionally users need the longitude and latitude or land area of census tracts or blocks for the computation of a summary

measure such as a residential segregation index. However, users can't select these items, or other items such as consolidated city population size code, place class code, or place description code from this section.

DADS works best when users are making a request for a small number of tables for a small number of geographic units. For example, this system works well for a user who wants to know the population size (a single cell) for all counties in Michigan or the distribution of income in Houston, Texas. However, many analysts need perhaps 10 or 12 tables (which might mean 200 or even 1,000 cells) for all zip codes in the nation. To get these data from DADS in its current configuration, one must list all the relevant zip code(s). Typing in over 10,000 zip codes is not a very practical alternative! In a typical STF request based on data stored on-line, one would select the appropriate summary level for zip code (820) and would get all the tables for all zip codes with the execution of one job. The configuration for block groups and tracts is similar, but one only has to highlight the tracts or block groups rather than typing them out. DADS can handle these requests for summary statistics for all zip codes or census tracks in the U.S., but it is a very labor-intensive task for the requester. The analyst who makes this sort of request is not just mining data that are never analyzed. The analyst is reducing 31,000 columns of information into 200-1,000 columns and then making the request for a unit of analysis that might range from around 3,000 for counties to more than 100,000 for block groups.

I'm certain that the future DADs will allow access to all summary tape file information. However, so far, only summary tape files 1 and 3 were released in CD-ROM form. Thus, none of the race-specific tabulations from STF2 and STF4 are available under the current Data Access system.

One would hope that DADS would allow the Census Bureau to eliminate the distinction between summary tape files and public use microdata files. It would be very useful for researchers to be able to define their own tables rather than being restricted to the limited number of tables supplied by the Census Bureau in its summary tape files. We had some researchers at Michigan recently wanted to look at disability according to race and sex. However, our researchers needed an age breakdown other than the typical 18-64 and 65+ groupings. Because the table they needed was not available in an STF file, they created one using PUMs files. Using PUMs, however, gave them a smaller case base, and the census geography could not be perfectly duplicated. In general, if analysts make a table or summary statistic based on a small number of variables, they should be able to get the tabulation for any level of geography. However, if they want to use 15 variables to define a summary statistic, then the level of geography becomes much more restricted (state, MSA, or PUMA). Thus, another advantage of eliminating the distinction between summary tape files and public use microdata files would be the increased sample size for the public use microdata files. Small populations such as male clerical workers, female pilots, 50 to 54 year-old women with own children under 5 years of age, or persons born in Guam could be studied better with a 16% count rather than the 5% files available with public use microdata. (On the other hand, I shudder to think that some of our researchers would be tempted to try to swallow the 16% count of white prime-age males when even the 5% count proves to be fairly cumbersome.)

FTP Access and/or Improving the Extraction Engine

Researchers who have excellent computing facilities may not want to get in the DADS queue every time they want access to census data. If the demand for the system is great, the Census Bureau may want to allow for FTP access so that users who have the capacity can bypass DADS except for quick exploration and for FTP access to the original raw files. If for reasons of confidentiality, the Census Bureau cannot provide access to the raw files—perhaps because confidentiality is built into the DADS software rather than into the data (via sample size or census geography)—then the DADS system needs to make improvements to the existing extraction engine.

The systems developed by CIESIN (Ulysses) and Public Data Queries (Explore) are extremely fast. One of the reasons they are so fast is that they produce is tables instead of the cases and variables used to produce the tables (or summary statistics). Any time one writes out individual records rather than tables or summary statistics, response time slows precipitously. If many users want micro-level extracts, as opposed to exploratory tables or even output from a summary tape file request (which is always a good example of data reduction), the response time will begin to discourage and irritate users. If a user has the capacity to handle the raw files, the Census Bureau should allow the user to do so, and thereby free up time for users who need the CPU.

The creators of the Integrated Public Use Microdata Samples (IPUMS) have found that their data cannot be used by all who might be interested in them, partly because of the sheer size of the files (125G) and partly because their primary audience (historians) traditionally has had limited access to powerful workstations. Thus, the IPUMS creators developed an extraction system that allows a somewhat disenfranchised user to create a work file. (These users are not completely disenfranchised as they do have access to the internet.) However, response time will not be quick with the IPUMS data extraction system

because, at least for the short run, all extracts will be executed on a single Sparc20. Clearly, a user with access to large disk storage and a powerful processor will be better off running the extract at his/her own desktop. Of course, the calculus needed to figure out whether the extract should be executed by the IPUMS workstation or a local workstation is complicated by the fact that other products, such as an extract codebook and SPSS cards are created along with the IPUMS-based extract.

Researchers at my site, the Population Studies Center at the University of Michigan, have made countless extracts since the release of the 1990 PUMs files using an in-house program that rectangularizes the hierachical structure of these files. Turnaround time is relatively quick (45 minutes - 3 hours) depending on the sample being used (1%, 3%, 5%, 8%), number of states requested, the size of the file being written out, and the load on the system. We purchased most of the microdata from the Census Bureau for $4,800 (5% - $4,000 and 1% - $800). I don't have a count of the number of extracts performed over the past few years but conservatively it has been 500 which works out to be $10 an extract and is more likely to be over a 1000. DADS will not be able to provide this quick and cost-effective system for our users although many users will be ecstatic about the system that DADS will provide.

Improving the Extraction Engine

Researchers often need access to more information than the tabular data provided through the STF data extraction system. The need for exploratory analysis can be met with tabular data and summary statistics; and more time spent exploring data before analysis often means less information actually ends up being extracted because the user has a much better idea of what is needed for the actual analysis. However, researchers often need access to microdata so that they can estimate equations. The systems developed by CIESIN (Ulysses) and Public Data Queries (Explore) allow researchers and policy analysts to get means and crosstabs from PUMs data in a matter of seconds; but not all statistical needs can be met with simple means and crosstabulations. The Census Bureau is aware of all of this and provides a Data Extraction engine for microdata. However, in the case of hierarchical files similar to census microdata (CPS), the interface for extraction is quite awkward. The interface needs to be improved, particularly, if for reasons of confidentiality, access to microdata is limited to the Census Bureau extraction engine.

Currently, the extraction procedure requires users to extract the records separately by record type (household, family, person) even though almost all users want a rectangular product. Although, the user certainly can merge the household, family, and person records to create a rectangular file there does not seem to be a rationale for adding this extra step to the procedure. In addtion, merging across record types increases the possibility that a novice user will end up with an erroneous file and not realize it. Novice users would also benefit from features such as variables that provide counts across the household, (e.g. the number of children under 4 or the number of earners) and the option to rectangularize the record based on something other than record type (e.g., rectangularize by household relationship for husband/wife or a mother/child file). Another common request is to select all person records if any person in a household meets a certain criteria, such as foreign birth, unemployment, age 60+, or interstate migration. Of course, the more bells and whistles that are added to the extraction engine, the more likely it is that people will use it for data management rather than just data access.

Archival Issues

The final question that a system like DADS invokes is how it can be archived. How will the Census Bureau unpack DADS so that they can turn over raw data and a codebook to the National Archives? Will there even be a codebook if the Census Bureau intends not to disseminate raw files and technical documentation as it did in the past? If the confidentiality firewall is built into the software, how can this information become part of the raw data so that confidentiality requirements continue to be fulfilled? Much of DADS sounds dynamic, which suggests that the system will be updated to include more data and perhaps that variables will be recoded to meet the demands of users. At what point will DADS be stabilized so that there is an archival record?

Table 1 has a list of questions that can help provoke our thinking and serve as guidelines in determining who should be responsible for making an archive out of DADS. Given the complexity and enormity of DADS, we may be tempted to allow the Census Bureau to be the archive for the census of 2000 and for all future data products, particularly because the Census Bureau looks like the archival expert when compared to the National Archives on many of these questions. However, it is important to remember that most state and federal data producers have poor long-term memories about old data (sometimes the definition of old is just a few years) and that there has been a lack of institutional memory within the Census Bureau about previous data losses due to poor archival policy. An article by Dollar nicely summarizes the historical record of federal data producers and the National Archives. In the decision on whether to archive summary statistics versus microdata from the 1940 census the reasoning was "if the Government agency that created the records for statistical purposes did not fully exploit them, it is hardly likely that anyone else will." (Dollar, 198x: 79). Thus, 1940 microdata were expendable.

Table 1

Who Should be Responsible for Data

(1)Is there expertise in the creating agency that can explain the context, technicalities of the subject area, or the idiosyncrasies of the data which would not be available if the records were transferred to an archive? Will that expertise remain available for all electronic records, or only for those in active systems.

(2)What functionality of the system used to create the records is necessary to meet the needs of archival users? Can the archives provide the necessary degree of functionality, or is the creating agency the only economically or technologically feasible place to preserve the data in a usable format?

(3)Will the creating agency guarantee equitable access within freedom of information and confidentiality policy guidelines?

(4)Do the records have continuing value to the creating agency so that it has an interest in and need to maintain the records beyond an external requirement?

(5)Will there be a duplication of effort if the archives acquire electronic records that have continuing value to the originating office?

(6)Where will the risk of loss or destruction be minimized?

(7)Can the creating agency guarantee that it will stabilize and not alter the archival record?

(8)Do regulations prohibit transfer of records from the custody of the original agency?

(9)What is the total cost to the organization to maintain electronic records for accountability and research purposes? How can these costs be reduced for the institution as a whole, without eliminating services to users?

Source: Hedstrom, Margaret. 1991. "Archives: To Be or Not to Be: A Commentary." Archives and Museum Informatics, Technical Report, Number 13.

References

Dollar, Charles. 1979. "Machine-Readable Records of the Federal Government the National Archives.", *Archivists and Machine-Readable Records: Proceedings of the Conference on Archival Management of Machine-Readable Records.* Edited by Carolyn G. Geda, Erik W. Austin, and Francis X. Blouin, Jr.

Hedstrom, Margaret. 1991. "Archives: To Be or Not to Be: A Commentary." *Archives and Museum Informatics, Technical Report*, Number 13.

1.Paper presented at the annual meetings of IASSIST, May 15, 1996, Minneapolis, Minnesota.

Social Science Data Services During the Last Five Years of the Millennium: Developments in the Delivery and Support of Data Services for Academic Research in Europe and North America.

by Adam Lubanski[1],
Information Systems Manager
ESRC Centre for Economic Performance

Introduction, Aims and Background

In general, data librarians are supported by researchers and computer staff in their view that demands for data services are likely to increase. Even where there have been technology related savings[2], other technology based tasks have arisen to add to the number of tasks performed by data support services - for example, the management and/or construction and maintenance of Web interfaces to data and associated documentation and literature.

The case for investing in data support services may seem clear to members of organisations such as IASSIST, CSS and Cause. And in the more recent reports produced by research/teaching support funding bodies such as the UK's ESRC and JISC there has been a marked increase in references to the data support environment. The main aim of this paper is to see if there is some empirical basis for the claim that investment in the continued development of data support services is worthwhile. The establishment of this claim will provide a sound basis from which to present the likely development scenarios of academic data services up to 2000.

The background to this study is associated with observations of a number of trends in institutional policy in the broad areas of social science support and administration in European and North American universities and associated research centres. These trends include:

- increasing funding pressures on researchers and research supervisors to speed-up submission rates - for example, from 1998, the UK's Economic and Social Research Council, (ESRC) will only fund PhDs at institutions where 60% or more students submit their PhDs within 4 years (currently, this is set at 50%)[3];
- the development of institutional measures to ensure researchers (and teachers) have necessary data resources and appropriate information systems (IS) infrastructure - about 70% of US academic institutions claim to have IS strategies[4]; and
- changes in IS infrastructure which have affected the resource demarcation between hitherto autonomous entities - for example, the integration of some or all of audio-visual, computer, data, library, network and telecommunications services[5].

To this can be added societal changes, such as the rise of so-called *meritocratic* practices such that "good policy" requires that position/status and associated resourcing have some empirical basis, as in *evidence-based planning* requirements[6].

The first set of facts gathered in this study relates to the current and projected growth rates in empirical research. Whichever way these are indicated, this growth is dramatic. The results of three methods of assessing empirical trends are summarised as follows:

- article-content analysis shows a consistent growth in the proportion of empirically based journal articles (Oswald, 1992; Figlio, 1994; Stigler, 1995; and Platt, 1996)[7];
- data access enhancements, particularly those associated with networking and interfacing, continue to speed-up the process of acquiring data and associated bibliographic references - for example, BIRON, BLS, IBSS, ICPSR and many local developments such as the data subsetting services at the NBER, SSDC and CEPIS[8]; and
- IT enhancements (storage and processing) have enabled major increases in productivity - recorded in studies of empirical research outputs (CEP/LSE) and business productivity measurement (MIT's CCS)[9].

The Fulbright Study is the basis for the second part of this investigation. It captures data support experiences from three perspectives: research, data services and IT/computer support. On the issue of efficiency of research and in-house data support, views are summarised as follows:

- the researcher-teacher view (28/30) is that data support (local and central) is an essential component of an efficient research environment - but, according to some (10/30), this may not be for ever;

- the data support service/person view (25/27) is that data and information services are experiencing a major upturn in demand - from both research and teaching activities (4 respondents also cited an increase in administration demands for data advice); and
- the majority IT/computer support service/person view (17/26) is that the acquisition of information and data support skills has become vital to their career prospects - others felt that networking and teaching support together with some integration of audio-visual support appeared a more fruitful path.

What seems obvious to the stakeholders, however, may not be fully recognised by the funders and planners. Ultimately, *in the long run*, data support funding will be determined by economic criteria.

The bad news at the present time, is that many data services are not well placed within the *order of things* to ensure that their strong economic arguments are well represented.

The good news is the data.

Background Data on Data Support Services

The selection of ninety or so interviewees, split about equally by the above types, was based on publication-citation methods (Gutman Library, February 1995). Briefly, this method adopted the following research sequence:

Data was captured mainly during 30-40 minute interviews (during some thirty visits to North American research institutions, March to May 1995); additional data was gathered from preliminary Internet searches and follow-up email to interviewees - typically, clarification of interview notes.

Supplementary *environmental evidence* was gleaned from institutional policy documents - as they related to data services/ support, and collected during the study. These included institutional responses from LSE, ESRC, national and state archives, data suppliers (e.g. the BLS) and a sample of North American universities.

The Sample Population:

Interviewees	Researcher	Data Support	IT Support
Total interviewed	30	27	26
Female[1]	6	18	5
Male	24	9	21
Job	9 SRAs (Senior Research Assistant/Officer	8 Data Librarians 3 Data Archivists	7 IT Assistants
	13 Professors	3 Data Consultants	14 IT Managers
	8 Directors (i.e. Directors of Research Centers	2 Info. Managers 4 Res. Managers 7 Data Managers	5 IT Directors
Research Center = 8	8 (of 400 fte)	8 (of 20 fte)	8 (of 14 fte + IT)
(Data Center = 7)	5 (of 100 fte)	7 (of 22 fte)	7 (of 15 fte)
University = 20 (17 Research)	17 (of ? fte)	12 (of 25 fte)	11 (of ? fte)
Published	27 (IBSS)	15 (Cause/effect, IASSIST, IBSS)	8+ (Cause/effect IASSIST, IBSS)
Cited	23 (ISI) 90? citations	12 (Cause/effect, IASSIST - approx.	Not counted
(Size/scale - average)			
Research Center	80 fte	2.5 fte	2 fte
Data Center	50 fte	2 fte	3 fte
University - research	5,000s+?f	2 fte	30 fte (LSE)
University - teaching	6,000s+?f	1 fte	30 fte (LSE)

1. An empirical basis for the prominence of women in computer-based data support is reported in Anderson, R.E., 1987.
? Denotes that figures were not noted at time of interview
(All figures are for economic/social sciences. IT Support includes networking and systems staff.)

Of seventeen research universities visited, virtually all (16) had a level of local data support far higher than that found (informally) in the UK. The one university which did not claim to have any formal arrangements for data support did, however, provide a very competent IT and computing advisory service together with a catalogued tape library facility. Basic advice about dataset management tasks was given by a Program Advisory team which referred detailed queries to an "analyst programmer with experience of databases".

Of the sixteen research universities claiming to provide a "resourced data service", four were classified as having basic data support[10] ; nine were classified as having intermediate data support[11]; and three fitted the classification of full data support[12].

Level of Data Support	Basic Data Service	Intermediate Data Service	Full Data Service
Research Universities (17)	4	9	3
Research Centers (8)	1	3	4

- of all forty-eight respondents interviewed at the 17 research universities, 45 expressed unprovoked favourable opinions of data support services - although nearly all said this was an under resourced area (and were actively lobbying for better funding); only two researchers said that their level of data support was adequate for local needs; and
- about half of the data support services/libraries were managed by the library service - a trend which was generally welcomed, but opposed by 4 respondents (3c and 1d - i.e. three computer staff and one data support person) who favoured independence.

Of seven research universities with large independently funded economic/social research centers - i.e. similarly configured to LSE and CEP:

- all seven had data support facilities (often named data libraries) both centrally based - typically, managed by the library (1) or IT services divisions (2), sometimes independent (4) - and devolved in the research centers themselves (data from interviews held at Harvard/NBER, Princeton/OPR, Cornell/CISER, Syracuse/CPR, Wisconsin/SSC, Ohio/NLSY and Michigan/ICPSR);
- both central and devolved models of data support appeared to function and coordinate well (according to interviewees), and were associated with high levels of researcher (and support staff) satisfaction; and
- all seven researchers (all experienced professors) interviewed had recently visited research universities in the UK (typically, the LSE and one or two others), and expressed some dismay at the poor level of data and IT support facilities for researchers - although conventional UK academic library facilities were rated highly.

Teaching universities provided data services through the library and IT/computer services. The VP of one university described an "innovative plan" for creating information (and data) support teams attached to academic departments and managed by the Library Service. Each team would comprise a "subject librarian", a "computer/network adviser" and an "AV-graphics-teaching resources manager".

Overall, although a few data staff had major reservations, this sort of reorganisation - the CLIO Model - was expected to be a feature of the IS future in teaching institutions. Of nine such support staff interviewed, all looked forward to re-defined jobs, some with enthusiasm (5) others with apprehension (4).

Resources and costs associated with data support

Facilities (IT infrastructure)

All respondents seemed aware of the major time-savings enabled by technological advances. In particular, researchers were keen to cite benchmarks for various modelling and statistical tasks. The following examples are typical of a dozen or so proffered. These were provided by an Industrial Relations researcher (LSE and DTI) and a trade/productivity research economist at ESRC's CEP:

Year - Stata v3	Machine	Cost (new)	Time (approx.)
1985	PC-XT	$=£1,400	6,000 secs
1990	386DX20	$=£2,000	500 secs
1993	486DX33	$=£2,000	90 secs
1995	Pentium90	$=£2,000	23 secs
1996	PentiumPro200	$=£3,000	7 secs
1996	Sun Model20-71	$=£7,000	30 secs*

** Sun will run two identical jobs in less than twice this time (actually, 51 secs).*

The following times correspond to the running time of the same Gauss program solving a non-linear equation system for a grid of points.

Year - Gauss	Machine	Cost (new)	Time (approx.)
1993	486DX66	$=£2,200	68 secs
1995	Pentium90	$=£2,000	19 secs
1996	Dell Latitude (laptop P120)	$=£2,600	17 secs
1996	1996 PentiumPro	$=£3,000	7 secs
1996	Sun Model20-60	$=£7,000	20 secs*

** Unix times cannot be guaranteed if multi-user.*

Typical hardware platforms found in the research centers included several Unix boxes (HP, IBM and Sun) and about one 486/Pentium per fte researcher (excluding part-time postgraduates who typically shared pooled 386/486 facilities).

- Novell (stable), NT (expanding) and Unix (stable) network servers were typical - 486/Pentium servers (1 Gb to 5Gb) and Unix cluster (5 to 50Gb) were typical storage capacities
- A typical mid-sized research center had 52 dos/windows PCS, 10 Macs (Classic), 3 Unix, 1 VMS and 2 Novell servers - supporting about 200 postgraduate students and 40 fte research staff
- use of campus-wide email (with approx. allocation of 1Mb space per user) was typical in research centers - as opposed to own email installation
- most common/popular packages were WordPerfect/Word, Netscape/Mosaic, ELM/Pine/cc:Mail, Gauss, Excel/123/ QuattroPro, SAS/SPSS/Stata - little evidence of the use of programming languages such as FORTRAN and C.

Most interviewees reported major changes in the pattern of IT/computer support. For example, the central Program Advisory Service, still prevalent in many UK universities, had all but disappeared in the US institutions in the Fulbright Study. Typically, programmers had been relocated and redesignated as departmental or cluster IT support staff. In the department, experienced programmers were often expected to provide a wide range of skills, covering software and hardware installation as well as teaching support duties. Some common responses to this major structural change were as follows:

The majority of "ex analyst-programmers" experienced what they saw as a "deskilling process" - a minority were optimistic about the challenge/value of learning new skills. The majority of this group expressed disquiet over "cost recovery" policies, and some expected this to lead to their extinction.

Some data staff felt that lone researchers in particular had lost a valuable resource, the program advisor. Nearly all data staff said they now found it necessary to provide some programming support for basic data management tasks - typically, SAS, SPSS and Stata. Just over half of all data staff interviewed (i.e. 14 of 27) appeared familiar with one or other of these programs - most of these said they had always seen data management programming as part of their remit, although they also reported less demand for detailed program advice.

In turn, nearly all IT support staff (23 of 26) reported concern that their skills needed to be upgraded (19), or had already been upgraded (4), to cope with new information and data management tasks. Many computer staff reported a major decline in the demand for their programming skills, and some said they had stopped all programming activity "many years ago".

The following IT/computer issues were cited frequently:

- around one third of data staff stressed that "computer skills were a basic requirement for data support staff" (10d=10 citations)
- some IT support staff (below managers) were concerned that IT managers were not offering appropriate/relevant training for IT staff, particularly data skills (5c)
- the use of public PC facilities by students was often 100% with queuing at peak times, indicating that demand exceeded supply - although some universities experienced a decrease in use of public facilities, as students stayed off-site (long journeys, bad weather, good support for modem links or local networking, etc. encouraged purchase of laptops)

The 1997-2000 IT Outlook

As predicted by Richard Rockwell (IASSIST, 1993), more powerful personal desktop PC-Workstations (running Unix and Windows NT/95), have continued to enable researchers to process large-scale datasets extracted from local and wide area networks. All respondents in the Fulbright Study expected continued performance improvements in desktop processing and data management, enabling further gains in research output. There continues to be general optimism about the contribution of IT hardware and software advances and their contribution to greater research productivity. The demand for IT support of remote laptop and home computing (distance learning) is expected to increase, stimulated by the growth in quality teaching software. In the light of so much concern expressed about reorganisation of support services, we may expect a professional review of all research support services. IT, library and data support staff will take the initiative (data piloted) and produce a more user-oriented information service.

Data Consultancy/Services (local data support)
According to researchers and data staff, the delivery of data support has become far more proactive. All data staff provided examples of "going out there" to find datasets, to advise on the best use of the data service (and other larger data facilities) and to help construct enhanced services through interlinked Web pages.

Library based data staff were most enthusiastic about the contribution of CD-ROM based datasets; some others, particularly experienced data support staff, seemed sceptical about the ultimate value of this form of data dissemination.

Researchers, closely followed by everyone else, were perplexed by the management and demarcation CD-ROM data (typically supported by the library service) and data on other media (typically found in data centers). This was generally put down to some form of *historical determinism*, and there was little evidence of plans for change in this respect!

Researchers and data staff reported that Internet type enhancements to data services had become expensive to maintain. Expectations were high, following the early lead taken (voluntarily) by data staff in constructing useable interfaces to datasets.

Web weavers reported time costs between 2 hours and two days per week for basic to comprehensive coverage of data services. Much of this work had been undertaken without additional funding. Data and research staff had become *de facto* Web Advisors.

Invariably, data support staff expressed "grave concerns" about data security and quality - particularly, in environments of decreased IT support services.

The following data issues were cited frequently:

- researchers in research groups/centers appeared less interested in programming support - although lone researchers still

needed help (4d+1r=5 citations)
• researchers seemed more concerned about quality of data accessibility, particularly with respect to speed (21r+10d+6c=31 citations)
• data staff and some IT staff were troubled by the ease of passing on large-scale undocumented (or poorly documented) datasets (22d+12c=34 citations)
• a small number of experienced researchers were concerned about a possible decline in the quality of data analysis - due to the trend to increase in accessibility (2r)
• European data could be difficult to locate, and often impossible to acquire (8d+9r= 17 citations)
• the majority of researchers prefer to download data directly to their personal machines, using their own data checking skills (18r)
• some data bureau seem reluctant to develop user services - so ICPSR (good at data checking) were playing an essential role (5r)
• the majority of researchers preferred to download entire file - rather make "front-end decisions" - even in the case of very large datasets (21r)
• researchers were keen to support the central university data repository - saying good local availability was important to research productivity (23r)
• about half of the experienced researchers interviewed said they liked to send their research assistants to the IT/data center - the other half tended to seek assistance directly from IT and data staff as appropriate

The 1997-2000 Data Outlook

General expectation of increased researcher self-sufficiency (with network infrastructure and data support) in programming/computing tasks. Alongside this, more effort in enabling access to datasets through the Internet. Later rather than sooner (evidence suggests) someone brave will pull CD-ROM data together with other data media. Expectation, ***in the long run***, of large investment in distributed data services via Web/Internet - economists/accountants will work with data staff, network/communications staff and higher education planners to produce properly resourced infrastructure. In the mean time, data staff will continue to produce prototype Web Data Servers without proper funding, and to experiment with the linking of datasets, documentation and bibliographic information. Many data staff will change from being *de facto* Web Advisors to *de jure* Information Managers.

Data Archives and Data Services

• a significant number of data support staff (and others) were concerned about small-scale institutions - in particular, their inability to manage and afford big datasets (7d+5r+4c)
• even in larger institutions, data staff said that if funding problems persist, central archives such as ICPSR would become still more important (3d)
• a few IT support staff and researchers stated their preference for getting data directly from central large-scale/national archiving (Essex, ICPSR, Roper, etc.) which might assure the quality and security of important datasets (4c+2r)
• some experienced researchers appeared keen to get data direct from source, and to bypass both local and national data services and archives (6r)
• researchers and data staff based in specialist research centers expected to play a major role as data resource centers, claiming the "full set of research, data and computing expertise" at the necessary level of expertise to advise specialist research projects (6 of 7r + 7 of 7d + 5 of 7c)

The 1997-2000 Data Archive Outlook

There was some expectation of devolution of large-scale data archives - major research universities and research centers will negotiate to get data associated with their specialism direct from source. Specialist research centers will work with major archives to distribute datasets, associated materials and expert advice to high level research projects. Data archives will continue to distribute datasets to the majority of non-specialist institutions, and to provide some further "one-stop-shop" support for institutions unable to resource a local data service. Data archives will combine with national data services and social science information gateways to lead the management and coordination of specialised data services and

associated expertise based at universities and research centers. At an international level, they will plan and manage the network of specialist data servers, and they will jointly work towards making national datasets statistically comparable.

Structural/organisational trends

Whilst the majority of researchers and data staff supported developments in the integration of data services and libraries, some had major reservations - citing loss of autonomy, deskilling and reduced service as likely outcomes.

IT support devolution and cost recovery continued to alarm support staff and over-exercise IS managers. It appears that every institution has completed or is considering a major reshaping of teaching and research support services.

Most researchers (17) supported the development of "one-stop-shopping" - i.e. the integration of IT and data services - although some (4), typically experienced, researchers questioned whether extending the ranges of skills might dilute the expertise. One experienced researcher and a few (4) data staff viewed integration plans as "cosmetic", and counter-productive in that experienced data staff were likely to be lost or become disaffected in the transition.

Integration of data and library catalogues was fully supported. About half of data staff reported that datasets and library records "have been or are being fully integrated"; a third said it was "being planned"; and the others said they "expected integration of catalogues to happen soon".

Most researcher-teachers (18 of 21 interviewed) supported "the trend" to deliver research (project) based courses to undergraduates using "real datasets".

Some "research-led teachers" discussed the need for a more effective Information Systems structure to enable appropriate support for courses which required a range of data and information inputs together with more advanced information management and processing techniques - cf. courses which employ artificial intelligence methods[13]. In this scenario the popularity of the one-stop-shop was very evident.

Network and Communications remain central services - albeit with evidence of growth in the number of local Servers. The move towards full integration of voice and network services continues - reaching over 50% in research universities[14]. About half of the institutions visited charged for ethernet (or token ring) connections, and some added a rental charge - $130 for network installation (+ $5 additional annual rental in some) was typical.

There were reports of an increase in (hitherto flatish) demand for remote computing which IT support staff expected to further stretch their reduced (in real terms) resources. Teachers, students and researchers are already expecting computer advice from remote locations - i.e. from home, conference locations, etc).

The 1997-2000 Structural Outlook

The continued devolution of large-scale central IT services seems likely, although a few of the very successful central systems should be able to construct a professional/economic case. Professional groups such as IASSIST and CSS will cooperate to produce documentation of "models of successful research support systems". Joint work on teaching support will enable teachers to deliver remote/distant learning courses using real datasets extracted from central archives (for general/introductory courses) and specialist data servers (for advanced courses).

Problem areas

Data staff were seriously concerned over a number of data security issues. All experienced staff (23 of 27) said they had initiated (8 of 23) or were initiating (10 of 23) or would/should initiate (5 of 23) procedures for data checking in light of bad dataset transfers.

Large scale data transfers using FTP were commonly cited as error prone[15], and bad Windows transfers (via File Manager, particularly from CD-ROM) and tape backups were also reported.

The following problems were cited frequently:

• Proliferation of forms (8d)
• a few staff mentioned the importance of getting away from the "format statement", which was seen as problematic for researchers and time consuming for support staff (3d+1c+2r))
• variable extraction via Web interfaces was generally supported - but there were some fears that speedy extraction would mean misuse of data, particularly if "data alerts" were not built into the system (5d)
• researchers complained about the work overload at ICPSR which meant they had to plan for up to eight weeks delay from data order (via ICPSR and local Data Library) - one researcher said she advised colleagues to "order data on the expectation that you may need it!" (6r)

The 1997-2000 Problems Outlook

> Expect "contents to check contents", i.e. auto check for data consistency. Data users to feedback errors though speedy "feedback system". Overdue replacement of paper forms by electronic forms. Data staff will make a professional case for greater investment in the integration of metadata with datasets, and experienced researchers will advise Web Data Server designers on the attachment of appropriate data documentation to subsets.

Success factors and performance indicators associated with data support

All researchers interviewed showed a keen awareness of research technologies and their contribution to research productivity. About half said they were sceptical of Windows style GUIs, but all research respondents said that productivity gains from advances in operating systems (for example, multi-tasking and large memory management) had made a major contribution to their (and others) empirical research. About a third (9 of 30) volunteered detailed benchmark figures consistent with those cited earlier in this paper.

Most respondents said that the quality of output was higher due to both IT and data support (roughly equally, when prompted). Two respondents (senior/experienced researchers) said their own research benefitted very little from local data services and a great deal from IT support -system programmers advice. In one case, local data services did not feature at his institution - he tended to use highly skilled systems programmers to assist with data management tasks. Six researchers argued that their research could not be undertaken properly without assistance from local "highly skilled data support staff".

As might be expected, productivity issues cited by data staff invariably reflected the content of the recent IASSIST Newsletter/Journal coverage. The importance of documentation featured in all interviews. Content analysis of respondence to an open-ended question on "what matters most?" shows the following recent articles to be representative of the range of issues cited: for example, general issues covered in Rasmussen, 1995, on-line codebooks by Sheih 1995, quality and accessibility by Beedham 1995, production by Winstanley and standards by Greene 1995.

IT/computer support staff were much more likely than researchers and data staff to mention shortfall in training, both in terms of their own needs and the requirements of end-users.

Data staff were most concerned about getting additional resources for new developments such as the delivery of subsetting services and related documentation.

The Fulbright Study showed a strong association between high levels of local data support and good performance[16]. Nearly all researchers were keen to empathise with the data services view that local data services are a key factor in the production of highly cited research publications.

Unprompted, over half of all respondents expected data support to be a major component of developing teaching methods, particularly new and redesigned undergraduate courses.

There are also strong *a priori* grounds for associating data support with good research performance. The evidence for growth in quantity and quality of empirical research is very strong, and it is also clear that academics are rewarded for research performance measured by publications and citations.

The variables that most distinguished the academics in the sample who had been promoted from those who had not included rate of publication in refereed journals, level of citation, research grants applied for and obtained and the number of PhD students under a person's supervision. Likelihood of promotion was correlated negatively with self-reported commitment to teaching[17].

Expect the organisation of local research support to be investigated more rigorously with a view to expansion in light of its proven contribution to research and teaching productivity.

Some economic conclusions

One thing is for certain in this study: Researchers, data services managers and IT staff all feel the funding future to be uncertain. While they may have clear visions of the what the direction of research support services ought to be, they are nervous about policy-making.

The *bad news* is that this *concern is well-founded.* Most researchers and virtually all research support staff (outside the library) are badly placed *(in the order of things)* to make a big impact on resource policy. And, as virtually every interviewee in the study has mentioned, failure to compete professionally for the appropriate level of resources will not enable their data utopia to become reality - or even *virtual reality.*

The *quite good news* is that they do have lots of real data on the productivity benefits of data support to enable the construction of a strong case for further investment. They also have the skills to disseminate this evidence. What is required is a framework for evaluating the contribution, and for this they may need to find time to review the small but growing literature in *information economics.* Recent work on the *economics of the Internet* and on the *contribution of IT to business* may provide some clues as to what to look for and what to measure in the context of research inputs.

As in other areas, the returns to investment in research support services can be measured in terms of productivity changes, performance and consumer benefits.

A number of recent research papers claim that investment in IT is associated with increased productivity, increased consumer benefits but unchanged business performance[18]. According to Brynjolfsson (1993), these results are compatible with conventional economic theory, i.e. "... firms are making the IT investments necessary to maintain competitive parity but are not able to gain competitive advantage".

Productivity gains to business and benefits to consumers due to investment in IT have been found to be strong. However, the impact of IT on Business Performance seems to be slight, sometimes negative. It appears from a stream of IT literature on business performance (1989-1993, cited by Hitt) that firms are unable to increase their profits through IT investment; indeed, while IT may be creating enormous value, it may simultaneously be intensifying competition and enabling entry, and thus lowering prices.

The *really good news* for data support services is that their contribution to empirical research is *truly, widely and deeply* recognised.

It is time to invest some of the energy and enthusiasm of research and teaching support services into the production of an empirically based case for expansion.

References

URL References:

BIRON: ESRC's Data Archive system for data searching on-line at Essex University:
http://dawww.essex.ac.uk/biron.html

CCS: MIT's Center for Coordination Science at MIT:
http://ccs.mit.edu/CCSWP190.html

CEP: ESRC's Centre for Economic Performance based at the LSE,:
http://cep.lse.ac.uk/

IBSS: ESRC/JISC's LSE based International Bibliography of the Social Sciences at BIDS, via
http://www.niss.ac.uk/ or http://www.lse.ac.uk/ - restricted access

ICPSR: Inter-university Consortium for Political and Social Research:
http://www.icpsr.umich.edu/ICPSR_homepage.html

JISC (1996): Policy for JISC Dataset Services Provision at NISS, via:
http://www.niss.ac.uk/it/JISCdatapol.html

NBER: National Bureau for Economic Research PWT data subsetting service:
http://nber.harvard.edu/pwt56.html

SSDC: University of California at SD's Social Science Data Center,:
http://ssdc.ucsd.edu/ - restricted access to some dataset services

Bibliographic Notes:

Anderson, R.E. and Coover, E.R., 1976, "Academic social research organisations and computerization", *Social science information* 15 (4/5): 741-754.

Anderson, R.E., 1987, "Females surpass males in computer problem solving", *Journal of Educational Computing Research*, vol.3(1).

Brynjolfsson, E., 1993, "The Productivity Paradox of Information Technology: Review and Assessment.", Centre for Coordination Science, MIT.

Cave, M., Hanney, S., Kogan, M. and Trevett, G., 1989, *The Use of Performance Indicators in Higher Education*, Jessica Kingsley, London.

Centre for Economic Performance, 1993, unpublished *Report to ESRC Datasets Policy Committee*, prepared by Layard, R., Lubanski, A. and Wadsworth, J. on behalf of the Centre for Economic Performance, LSE.

Figlio, D., "Trends in the publication of empirical economics", *Journal of Economic Perspectives*, 8 (Summer 1994): 179-87.

Gaspar, J. and Galeser, E., 1996, Preliminary Draft, Communications technology and the future of cities, Stanford, Harvard University and NBER.

Hitt, L. and Brynjolfsson, E., 1995, "Productivity without profit? Three measures of information technology's value.", *MIS Quarterly*. (See CCS under URL References above).

Munson, J.R., Richter, R.L. and Zastrocky, M.R., 1994, Cause ID 1994 Profile.

Oswald, A. (1991) "Progress and microeconomic data", *The Economic Journal*, 101 (January 1991), 75-80.

Over, R., 1993, "Correlates of career advancement in Australian universities.", *Higher Education*, vol.26, no.3, pp.313-329

Platt, J., 1996, "Has funding made a difference to research methods?", *Sociological Research Online*, vol.1, no.1,<http://www.socresonline/1/1/5.html>.

Ramsden, P. (1994) "Describing and explaining research productivity", *Higher Education*, V.28, N2, p 207-226.

Ruus, L.G.M., 1990, "Planning a data service facility", *Data Library Service*, University of Toronto, 30/11/90 (internal document).

1.Paper presented at the annual meetings of IASSIST, May 15, 1996, Minneapolis

2 For example, in some areas the requirement for help with acquisition of data documentation has been reduced due to investment in on-line services.

3 ESRC Annual Report, 1994-95, December 1995.

4 From Cause ID Survey 1994. In the UK, the ESRC and JISC have recently produced their respective policies for dataset services - see URL References.

5 From Cause ID Survey 1994, supplemented by evidence from my own study - see Fulbright Report in URL References.

6 Verified from NCDS studies at University of Sussex, 1994.

7 Refer to Bibliographic Notes

8 Refer to URL References

9 Refer to bibliographic notes - Hitt and Brynjolfsson, 1995.

10 As defined by Laine Ruus' 'Planning a data service facility', in Ruus (1990).

11 op. cit.

12 op. cit.

13 For example, Richard Freeman's new economics course at Harvard University.

14 Cause ID 1994 Profile.

15 A CEP researcher reported 5% failure rate (identified in data checking) in FTP transfers of some 60 data files of between 5 and 25 megabytes.

16 As indicated by publication and citation rates. See Kogan, M., et al, 1991, in Bibliographic Notes below.

17 See Over, R., 1993, in Bib.

18 Hitt and Brynjolfsson, 1995.

INTERNATIONAL ASSOCIATION FOR SOCIAL SCIENCE INFORMATION SERVICE AND TECHNOLOGY

• • • •

ASSOCIATION INTERNATIONALE POUR LES SERVICES ET TECHNIQUES D'INFORMATION EN SCIENCES SOCIALES

The **International Association for Social Science Information Services and Technology** (IASSIST) is an international association of individuals who are engaged in the acquistion, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, data base librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompases hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the IASSIST QUARTERLY. They also benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:

Regular Membership. $40.00 per calendar year.

Student Membership: $20.00 per calendar year.

Institutional subcriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subcription:

$70.00 per calendar year (includes one volume of the Quarterly)

Membership form

I would like to become a member of IASSIST. Please see my choice below:

- ☐ $40 Regular Membership
- ☐ $20 Student Membership
- ☐ $70 Institutional Membership

My primary Interests are:

- ☐ Archive Services/Administration
- ☐ Data Processing
- ☐ Data Management
- ☐ Research Applications
- ☐ Other (specify)_____

Please make checks payable to IASSIST and Mail to :

Mr. Marty Pawlocki
Treasurer, IASSIST
% 303 GSLIS Building,
Social Science Data Archives, University of California, 405 Hilgard Avenue, Los Angeles, CA 90024-1484

Name / title

Institutional Affiliation

Mailing Address

City

Country / zip/ postal code / phone